Pichler, Heike. 2010. Methods in discourse variation analysis: Reflections on the way forward. *Journal of Sociolinguistics* 14(5): 581-608.

Methods in Discourse Variation Analysis:

Reflections on the Way Forward[1]

ABSTRACT

This paper demonstrates the need for a uniform model of discourse variation analysis which is equipped to capture the complex nature of discourse variation and change whilst also ensuring generalizability. A review of the literature shows that the current heterogeneity in corpus construction, data quantification and theorizing of discourse variables impedes reliability and intersubjectivity. Suggestions are offered to achieve comparability, and a case is made for a consistent integration of pragmatic function as a factor group in the analysis. The extension of the variationist paradigm to the level of discourse is discussed, and the need for a definition of discourse variables which caters for their flexibility and multifunctionality is demonstrated. It is argued that some methodological consistency is required in variationist discourse analysis in order to advance towards a holistic description of patterns in language variation and change which spans all components of the grammar, and to systematically explore how discourse features are used and manipulated to create social identities.

Key words: discourse variation and change, discourse variables, discourse markers, variationist methodology, generalizability

Running title: Methods in discourse variation analysis

Number of words in main text: 8,063

ABSTRACT

Dieser Beitrag plädiert für eine einheitliche Methode zur Diskursvariationsanalyse, welche der Komplexität von Dikursvariationen gerecht wird und Generalisierungen zulässt. Der Literaturüberblick ergibt, dass die gegenwärtige Heterogenität in der Zusammenstellung von Korpora, Datenquantifizierung und Konzeptualisierung von Diskursvariablen Zuverlässigkeit und Intersubjektivität beeinträchtigt. Es werden Vorschläge unterbreitet, wie man Vergleichbarkeit sicher stellen kann, und es wird für eine konsistente Integration von pragmatischen Funktionen als Faktorengrupppe in der Analyse plädiert. Die Anwendung eines variationistischen Ansatzes auf die Analyse von Diskursmarkern wird erörtert. Des Weiteren wird die Notwendigkeit illustriert, Diskursmarker als Variable zu definieren, wobei ihre Flexibilität und Multifunktionalität berücksichtigt werden muss. Methodische Konsistenz ist erforderlich, um eine umfassende Theorie von Sprachvariation und -wandel zu formulieren und um systematisch zu analysieren, wie diese Variablen zur Konstruktion von sozialen Identitäten verwendet werden.

Schlüsselwörter: Diskursmarker, Variation, Sprachwandel, Quantifizierung, Generalisierung

Titel: Methoden der quantitativen Analyse von Diskursmarkern

1. INTRODUCTION

The study of discourse-pragmatic features such as *oh, well, I mean, you know,* etc. has not traditionally fallen under the remit of variationist sociolinguistics. Because of their ill-defined grammatical status, their extra-sentential positioning and their lack of truth-functional semantic meaning, discourse-pragmatic features have for long been viewed as extra- or a-grammatical elements of language in traditional grammatical accounts.[2] Marginalized as overt manifestations of verbal dysfluencies and inarticulateness, meaningless verbal fillers and superfluous hesitation markers, they were dismissed as not warranting linguistic investigation. Scholarly interest in these features developed only in the 1980s, with the growing recognition that they carry social meaning, perform indispensable functions in social interaction, and constitute essential elements of sentence grammar (Dines 1980; Schiffrin 1987; Traugott 1995).[3]

Although variationist studies of discourse-pragmatic features are still relatively rare, especially when compared to the profusion of variationist studies of phonological and, to a lesser extent, lower-level morpho-syntactic features, they consistently highlight the fact that discourse features, like features in other components of the grammar, evince orderly heterogeneity and a capacity for change. It has been repeatedly shown, for example, that patterns of variation in terms of the frequency, strategic use and formal encoding of discourse features correlate with social and internal factors (e.g. Cheshire 1981; Holmes 1995; Macaulay 2005; Pichler 2009; Stubbe and Holmes 1995), and that these factors are motivating forces in discourse change (e.g. Ferrara and Bell 1995; Pichler 2008; Tagliamonte and Hudson 1999). Yet despite accumulative evidence that discourse features are systematically involved in patterns of language variation and change, it is fair to say that variationist study of discourse 'is still at an

elementary stage' (Macaulay 2002a: 298). In this paper, I demonstrate that the embryonic state of discourse variation analysis is due largely to the lack of a coherent set of methodological principles, and argue that the development of a reliable and uniform model of analysis is paramount in order to advance our understanding of discourse variation and change.

Quantitative studies of lower-level phonological and morpho-syntactic variation and change have been relatively homogeneous and congruent in focus and methodology, thus ensuring reliability and intersubjectivity, i.e., that the same results are obtained across different datasets and by different analysts (Bailey and Tillery 2004: 11). The resultant comparability (and, necessarily, the profusion) of studies has allowed scholars to formulate general principles as regards, for example, the role of women in language change and the relative importance of social and internal factors in language variation (Labov 1998; Preston 1991). Yet, these generalizations might not apply equally across all levels of the grammar for there is evidence that the mechanisms underlying variation and change in discourse (and higher-level syntax) are different from those in phonology and morpho-syntax (Cheshire, Kerswill and Williams 2005). Progression towards a holistic theory of language variation and change therefore necessitates that generalizations about patterns of variation and change in discourse (and syntax) be drawn independently. Alas, this endeavour is hampered by the persistent diversity of methods used in discourse variation analysis and the resultant lack of cross-corpora comparability. This makes it difficult, if not altogether impossible, to synthesize the results of previous studies into a set of coherent findings, and thus, to formulate empirically grounded generalizations of discourse variation and change. In this paper, I advocate methods which ensure reliable cross-corpora comparisons, and thus allow scholars to systematically explore the social and

internal dimensions of discourse variation and change as well as the role of discourse features in the marking and construction of social identities.

The diversity of methods in discourse variation analysis can be linked to the fact that discourse-level features 'are used by speakers in such complex and sophisticated ways [that] studying variation in their use is no straightforward task' (Stubbe and Holmes 1995: 85). Because of fundamental differences in the nature and use of lower-level and higher-level linguistic features, it remains a contentious issue whether the variationist paradigm, originally developed by Labov (1963, 1966) for the analysis of phonological variability, can be extended to the analysis of higher-level variability. Discourse-level features do not easily satisfy the criteria set out by Labov (1972) for the linguistic variable, the principal methodological tool in variationist analysis. Firstly, the operation of semantic and pragmatic constraints on discourse features affects the frequency critical for quantitative analysis. Secondly, their intrinsic multifunctionality prohibits discourse variants from being identified on the basis of semantic equivalence. Thirdly, their multifunctionality causes difficulties in circumscribing their variable context, which hampers attempts to quantify their variability in an accountable manner (see, however, D'Arcy 2005). These complexities have led to the adoption by discourse variationists of highly differential methods of analysis. The heterogeneity mainly affects the nature of the corpora investigated; the theoretical basis for defining discourse variables; the quantitative methods employed; and the optional integration of function as a factor group in the analysis.

This paper sets out to critically examine the methodological heterogeneity of discourse variation studies, with the aims of (a) assessing the reliability of the currently used methods, (b) identifying the parameters in which these methods

differ, (c) exploring the theoretical implications of these differences, and (d) proposing an analytical framework which is equipped to ensure reliability and intersubjectivity. I begin, in section 2, with exploring how comparability can be promoted despite the inevitable diversity in the construction of corpora used for discourse variation analyses. In section 3, I argue in favour of extending the linguistic variable to the analysis of discourse-level features, provided that it is modified to cater for their distinctive characteristics. Section 4 examines the respective suitability and value of different quantitative approaches to discourse variability – those which require a closed set of variants and those which do not. I posit in section 5 that in order to provide accurate accounts of discourse variation and change, it is necessary to integrate pragmatic function as a factor group in the analysis. Finally, in section 6, I provide a summary of the analytical methods advocated in this paper, outline what insights about discourse variation and change they may afford us, and explain how their adoption might benefit the variationist enterprise as a whole. In this section, I also discuss how a reliable and uniform framework of discourse variation analysis may inform efforts to explore how individuals manipulate discourse variation to negotiate and establish social meanings and identities.

I do not claim to answer here all the questions surrounding discourse variation methodology, but I hope to stimulate serious reflection and discussion amongst discourse variationists in order to generate a gradual progression towards the formulation of a coherent theory of discourse variation and change.

2. CORPUS CONSTRUCTION: CATERING FOR CONTEXTUAL CONSTRAINTS

The usage of discourse-pragmatic features is strongly constrained by the interactional and situational context of their occurrence, to the extent even that these factors may outrank the effect of social factors on discourse variability (see e.g. Schleef 2008). Table 1 lists quantitative studies into the effect of context on discourse-pragmatic variation, detailing which contextual factors affect the variables' use, and whether it is their frequency or their function that is affected. A wide array of contextual factors is involved: discourse type and activity context; topic, purpose of and attitudes to the interaction; speaker roles and relationships; and communicative channel. The fluctuating frequencies of discourse-pragmatic features are generally attributed to their communicative function(s) being more or less relevant across different contexts (see, for example, Freed and Greenwood 1996), while their differential functional uses across context are often attributed to the variable interactional demands of different speech events (see, for example, Mauranen 2004).

[insert Table 1]

The extreme context-sensitivity of discourse features hampers cross-corpora comparability and generalizability. With comparability not generally a design feature of dialectal corpora, different corpora may produce differential opportunities for the occurrence and strategic use of discourse features. It may therefore prove (near) impossible to disentangle the effect of contextual, social and geographical factors on the observed variation. Terraschke (2007), for example, failed to establish whether the differential rate of general extenders, i.e., clause- or utterance-final constructions such as *and that, and stuff like that, or something*, in her and Overstreet's (1999) corpora is due to dialectal differences (New Zealand vs. American English) or contextual differences in the relationship between interactants (strangers vs. familiars).

Difficulties in cross-corpora comparisons could potentially be resolved if scholars were to construct corpora on parallel principles. Yet this is not a realistic scenario. Firstly, corpora are collected with different goals and assumptions in mind, and researchers work with different resources. These factors control the overall design features of private corpora, yielding a wide range of vastly differential corpora. Secondly, as Kallen and Kirk (2007) highlight in their discussion of the construction of the ICE-Ireland corpus, even where scholars aim to construct comparable public corpora according to global design principles, they face local challenges that require some elaboration or modification of these principles. In the absence of consistent corpus construction principles, scholars therefore need to exercise great caution when comparing corpora of socially and regionally diverse speech samples, and consider at all times the possibility that cross-corpora differences may reflect the effect of differential contextual constraints on the variation, rather than actual social or geographical variation (see, for example, Escalera 2009).

The isolation of contextual factors from other factors can be facilitated by implementing the following measures. Firstly, more studies need to be conducted which test the effect of multifarious contextual factors on the usage and distribution of the range of discourse-pragmatic features. Secondly, detailed demographic as well as textual metadata about corpora need to be provided. In conjunction with detailed knowledge of contextual effects on the use of targeted discourse variables, knowledge of metadata will allow scholars to assess – maybe not with total but at least with some confidence – whether the findings obtained across corpora are comparable, or whether cross-corpora differences in the use and distribution of discourse variables are an artefact of differential contextual constraints. Terraschke's (2007) comparison of her data with Overstreet's (1999)

would yield more conclusive insights about dialectal variation in general extender usage if there was cogent evidence from other studies about the effect of interpersonal relationships on the usage of these constructions, and if detailed provision of metadata by both scholars allowed us to rule out the potential effect of any other contextual factors on cross-corpora variability. The type of metatextual data required for all studies of discourse variation and change is summarized in Table 2.

[insert Table 2]

3. DISCOURSE VARIABLES: DELIMITING THE VARIABLE CONTEXT

In order for discourse variation analysis to progress towards accountability, comparability and generalizability, it is crucial to address another issue which has not yet been satisfactorily resolved: whether discourse-pragmatic features can be operationalized as linguistic variables, and if so, on what grounds.

The concept of the linguistic variable was originally developed by Labov (1963, 1966) for the analysis of phonological variability: to fully understand the mechanisms underlying linguistic variation, it is necessary to isolate the whole set of possible variants of a variable and to report the total number of the variants' actual and potential occurrences ('Principle of Accountability,' Labov 1972: 71-72). Whilst Sankoff (1973: 44, 58) argues that the extension of the linguistic variable to levels 'above and beyond phonology' is 'not a conceptually difficult jump,' others have expressed their reservations about this extension.

Dines (1980) and Lavandera (1978) questioned the methodological and theoretical soundness of scholars' wholesale transfer of the linguistic variable to the analysis of non-phonological variability on the basis that the defining criterion of phonological variables, i.e., semantic equivalence, is not easily met by higher-

level linguistic variables due to their differential semiotic nature. Their proposals to modify the defining criterion of the linguistic variable for the analysis of discourse (and syntactic) features were criticised by Cheshire (1987) and Romaine (1984) for being made without consideration of the theoretical assumptions underlying the variable's original conception. Yet it is hard to conceive how the mechanisms underlying higher-level variability can be satisfactorily explained without modifying the linguistic variable in accordance with the properties of the grammatical level to which it is applied.

The notion of the discourse variable has to account for a very different phenomenon than notions conceptualized for variables in other components of the grammar, including that proposed by Serrano (to appear) for syntactic variables.[4] Unlike other linguistic features, discourse-pragmatic features do not occupy a fixed syntactic or segmental slot nor is their pragmatic meaning constant; they occur in a variety of positions and take on different pragmatic meanings across different contexts of use. They are referentially and syntactically optional elements of discourse that can be omitted without necessarily altering the propositional meaning or syntactic structure of an utterance. Consequently, their usage in discourse is triggered by different motives than that of other linguistic features; discourse features are not generally employed to communicate content but to express speaker attitudes and guide hearers' decoding of messages (see Section 5 below). A satisfactory definition of discourse variables, then, has to account for the unique nature of discourse-pragmatic features as well as the possibility that, due to their unique nature, variability in their use might be motivated by a different combination of factors than that of other variables. The big question is which unifying criterion to employ to establish equivalence relationships between discourse variants. It cannot be one of semantic equivalence

since discourse features are semantically bleached and thus, by definition, lack semantic meaning.

Dines (1980) and Lavandera (1978) propose to substitute the condition of semantic sameness for one of functional comparability. Dines illustrates this conceptualisation of discourse variables in her analysis of general extenders (or what she calls 'set-marking tags'): *and that, and stuff like that, or something,* etc. The equivalence relationship between the variants is established on the basis of their performing a common function in discourse, i.e., 'cue[ing] the listener to interpret the preceding element as an illustrative example of some more general case' (Dines 1980: 22). The function-based conceptualisation of discourse variables has been very influential and has been widely adopted in subsequent studies of discourse variation, particularly in studies of quotatives ('all strategies used to introduce reported speech, sounds, gesture and thought by self or other,' Buchstaller 2006: 5) (see also Macaulay 2001; Tagliamonte and Hudson 1999) and intensifiers ('every option speakers have at their disposition to reinforce or boost the property denoted by the head they modify,' Rickford et al. 2007: 7) (see also Ito and Tagliamonte 2003; Macaulay 2006). This conceptualisation is not without its problems, though.

Firstly, if we take function as the theoretical basis for defining discourse variables, it might sometimes be unfeasible to list variants exhaustively and to adhere to the principle of accountability. Function is an open category that spans different components of the grammar. Hence, identifying all variants of the targeted functional category and noting each variant's actual and potential occurrences might be beyond the remit of individual analyses. Also, to include a range of elements from different components of the grammar 'in the same analytic unit [...] is surely stretching the concept of the variable beyond all credibility'

(Cheshire, Kerswill and Williams 2005: 164). Yet unless we close the set of variants, we might not be able to fully explain the variation found in our data (see sections 4.2 and 4.3).

Secondly, we need to acknowledge that discourse-pragmatic features are polysemic elements, and that function is not a stable denominator. As pointed out above, Dines (1980: 23) argues that different surface forms of general extenders are related by virtue of their 'common function of marking the preceding element as a member of a set.' On this basis, the general extender system can be treated as a closed set, and the proportional distribution of variants can be modelled within a standard variationist framework (see, for example, Dubois 1992). However, this approach might oversimplify the nature of general extender usage. General extenders perform a wide range of functions beyond evoking a larger set (Aijmer 2002; Cheshire 2007; Overstreet 1999) and their uses are far from being restricted to set-marking functions (Cheshire 2007; Pichler and Levey under review; Tagliamonte and Derek 2010). Also, different forms of general extenders are associated with different discourse functions (Aijmer 2002; Overstreet 1999). These insights suggest that general extenders are not in fact a unitary functional category. In other words, they cannot be treated as an analytical whole on the theoretical basis of having a common discourse function.

This view receives additional support from the fact that in the process of grammaticalization, discourse-pragmatic features over time develop new functions in addition to, or instead of, their original semantic meanings (Brinton 1996; Traugott 1995). Yet even if the original definition of general extenders' core function was modified to reflect the semantic-pragmatic meaning shifts associated with their diachronic development, this would not satisfy variationists' needs. Firstly, if contemporary studies defined general extenders as performing a

different core function than that postulated in Dines (1980), they would strictly speaking no longer be analysing the same variable as Dines. Any comparisons between these datasets would therefore be questionable on theoretical grounds. Secondly, if scholars defined discourse variables on the basis of the functions they perform in their data, we might end up with diverse definitions of the same underlying feature. After all, the functional spectrum performed by discourse-pragmatic features is highly contingent on context (see section 2 above). Again, cross-corpora comparisons would not be theoretically defensible. What we need, then, is a conceptualisation of discourse variables which (a) encapsulates the different meanings which discourse-pragmatic features develop in the course of their grammaticalization (see also Schwenter and Torres Cacoullos 2008: 11-12); (b) allows us to capture discourse-pragmatic features' polysemic layering, that is, 'synchronic variation among different meanings in the same form' (Torres Cacoulos 2001: 462); and (c) allows us to quantify functional variation and change (see section 5). Definitions of discourse variables based on functional comparability are inadequate for these purposes.

A definition of discourse variables which is better equipped to capture the complexity of discourse variability is one based on its variants' underlying structural similarity. In her research into patterns of variation in the formal encoding of discourse-level elements, Pichler (2008, 2009) defines the discourse variables I DON'T KNOW and I DON'T THINK as constituting fixed multi-word constructions whose variants, though not necessarily sharing the same discourse function, are formed with the same components: the first person singular pronominal subject 'I', negative periphrastic DO, and the predicates 'know' or 'think'. Pichler extends this conceptualization to negative polarity tag questions, defined as identifiable syntactic constructions whose variants occur in the same

syntactic environment (appended to an utterance) and are formed with the same grammatical components (auxiliary, negative particle, pronoun). Most recently, structure-based conceptualizations of discourse variables have also been adopted in studies of general extenders, which have been defined on the basis of 'prototypically shar[ing] a structural pattern schematically represented as: (connector) (modifier) (generic noun/pro-form) (similative) (deictic), where parentheses indicate the optionality of individual components' (Pichler and Levey under review) (see also Tagliamonte and Denis 2010).[5]

Defining discourse variables on the basis of their underlying structure has several methodological and theoretical values. Firstly, this conceptualization caters for discourse-pragmatic features' diachronic meaning changes and synchronic polyvalence whilst still ensuring that '*the variants are in some way the same*, have something in common' (Dines 1980: 18, italics in the original). It thus enables scholars to draw cross-corpora comparisons of the distribution of a *form* which has developed different pragmatic meanings in different varieties under the envelope of one discourse variable. Secondly, because in this definition, discourse features are not reduced to some fabricated core meaning, it allows scholars to include function as a parameter in the quantitative analysis and investigate whether functional variation and change impact on the distribution of variants (see further Section 5). Thirdly, if the theoretical basis for delimiting discourse variables is structural equivalence rather than functional comparability, it is possible to close the set of variants that defines the variable, and to conduct an accountable quantitative analysis along the parameters of the variationist paradigm (see Section 4.1). How we conceptualise discourse variables thus has important implications in terms of the questions we can answer.

Some discourse variables might be better conceptualised based on functional comparability between variants (e.g. intensifiers), others based on structural commonality (e.g. general extenders).[6] Such differences are an inevitable result of the diverse formal and functional nature of the class of discourse-pragmatic features which makes it unfeasible to categorically apply one conceptualization to all of these features. What is important is that scholars are consistent in how they conceptualise specific discourse features and that they set out clearly how they have delimited the variable context. In the past, scholars have sometimes quantified variation in the use of discourse-pragmatic features without conceptualising them as discourse variables (e.g. Cheshire 2007; Ferrara 1997). Yet, unless scholars identify the forms they analyse under the umbrella of a variable, it may prove difficult to establish how the different forms included in the analysis are related to each other, what the theoretical basis is for grouping them together, and what forms are, in fact, included in the analysis. This problem can be circumvented by employing the functionally- or structurally-based discourse variable as a heuristic analytical device, thus making explicit which forms have been included in the analysis and what the commonality is between them.

4. DATA QUANTIFICATION: MULTIPLE REGRESSIONS, FREQUENCIES, ALTERNATIVES

Variationist sociolinguistics aims to 'express in quantitative terms the strength and association between a contextual feature and the linguistic variable' (Bayley 2002: 118). For discourse variability, three quantitative approaches to uncovering these correlations can be distinguished: one which is employed with features conceptualised as discourse variables with a closed set of variants; one which, in the absence of a closed universe of variants, relies on frequency tabulations to

reveal patterns of variation and change; and one which, also in the absence of a closed universe of variants, delimits the variable context syntactically.

4.1. Quantification of Variables with a Closed Set of Variants: Multivariate analyses

A key advantage of conceptualising discourse-pragmatic features as linguistic variables – whether on functional or structural grounds – is that it allows researchers to close the set of possible variants and report their proportional frequencies out of the variable, in line with the principle of accountability. Yet, because linguistic variation is generally constrained by multiple social and internal factors, overall frequencies and proportional frequencies of variants across independent variables alone do not fully explain the mechanisms of variability in the data nor do they detect evidence of linguistic change in every instance. In order to provide reliable descriptions of the social and internal conditioning of discourse variation and change, it is necessary to conduct a multivariate analysis of the data which can quantitatively assess the relative effect of multiple intersecting independent variables on the data when they are considered simultaneously. The method of multivariate analysis most commonly used in variationist sociolinguistics to date is variable rule analysis with Goldvarb X (Sankoff, Tagliamonte and Smith 2005).[7] The statistical modelling techniques of this method reveal three lines of evidence: (i) statistical significance of individual factors and factor groups, (ii) relative strength of factor groups, and (iii) constraint hierarchy within factor groups (Tagliamonte 2002: 733). They provide a precise and replicable measure of the patterns of variability and change in the data.

Despite its intrinsic values, multivariate analysis is not consistently applied across discourse variation studies. Many studies report the results of univariate

statistical analyses, i.e., they demonstrate that multiple social and internal factors constrain the occurrence of discourse variants without giving any indication of their relative importance (e.g. Andersen 2001; Macaulay 2005; Stubbe and Holmes 1995). Macaulay (2006), for example, reports the distribution of intensifier *pure* in Glasgow teenage speech across a range of social and internal factors (age, gender; predicative vs. attributive position, negative vs. positive evaluative purpose, syntactic position). Yet because he does not subject his data to multivariate analysis, we do not know which of these factors makes the most important contribution to the occurrence of *pure* in this variety, or which constraints might be shifting in importance across age groups. The study therefore does not fully capture the variable grammar.

Another drawback of univariate analyses of discourse variability is that they hamper cross-corpora comparability. Pichler's (2008) hypotheses about the differential social embedding of the grammaticalization of *innit* in the north and south of England has to remain tentative because Andersen's (2001) univariate analysis of *innit* in London English does not reveal the relative importance of different social factors when their impact is considered simultaneously. Only when multivariate methods are employed consistently across corpora (and variables and variable contexts are defined and delimited along the same parameters) can subtle differences in the significance, strength and ordering of conditioning effects reveal whether and to what extent different varieties share an underlying grammar (Tagliamonte 2002).

Waters (2009) conducts a multivariate analysis to compare the impact of five contextual factors (age, sex, education; polarity, position) on the use and distribution of *actually*, conceptualised as a co-variant of *really*, in Toronto and York English. Her analysis reveals interesting correlations indicative of the

following cross-variety differences: (i) *actually* is more grammaticalized in Toronto than York English; (ii) *actually* is a marker of social group membership in York but not in Toronto. In their studies of quotative BE + *like*, Buchstaller and D'Arcy (2009) and Tagliamonte and Hudson (1999) convincingly demonstrate the value of multivariate analysis for tracking the diffusion of innovative discourse features across national varieties of English. These studies illustrate what insights multivariate analyses afford us about cross-variety developments in discourse as well as the social and internal dimensions of discourse variation and change more generally.

4.2. Quantification of Variables without a Closed Set of Variants (1): Frequency Tabulations

Multivariate analysis necessitates that discourse variables be defined as a closed set of functionally or structurally comparable variants. Yet discourse features cannot always be conceived of in this way because it is not always clear what other forms might be their co-variants. In this scenario, alternative methods are required to compute discourse variability in an accountable and replicable manner.

Some scholars report raw frequency scores of discourse-pragmatic features in their attempts to reveal sociolinguistic differences in their use (e.g. Dines 1980; Dubois and Crouch 1975). For example, Erman (2001) reports on the basis of raw numbers that young speakers increasingly use the discourse feature *you know* for different pragmatic functions than adult speakers. Freed and Greenwood (1996) report on the basis of raw numbers that the frequency of *you know* is stable across gender and variable across contextual factors. Inevitably, though, individuals and social cohorts produce different amounts of speech and thus have differential opportunities for producing discourse variables.[8] The fact, then, that there is no indication in these studies that the comparisons are balanced in terms of the

amount of speech produced by individual cohorts and in different contexts casts some doubt on the validity of Erman's (2001) hypothesis that *you know* is undergoing functional change, and Freed and Greenwood's (1996) conclusion that its occurrence is constrained by contextual factors, not gender.

It is therefore crucial that scholars report *relative* frequencies of discourse features in order to yield reliable and, importantly, replicable results (see also Buchstaller 2009). The important question here is: relative to what? If we aim for utmost comparability, we require normalised quantification methods which produce the same results across diverse datasets. Methods such as counting frequencies per line of transcript (e.g. Vincent and Sankoff 1993) or per minute/hour of speech (e.g. Meyerhoff 1994; Siegel 2002) are problematic in this respect since the denominator – line of transcript, minute/hour of speech production – is not stable but variable in length, both across datasets and across individuals.

A seemingly more accountable method is to index raw scores of discourse variables as normalized proportions of total word counts (e.g. Andersen 2001; Cheshire 2007; Fuller 2003; Macaulay 2005; Stubbe and Holmes 1995). To compute this index, the total number of tokens of a discourse variable produced by an individual/group is divided by the total number of words produced by this individual/group. This count is then multiplied by 1,000 or 10,000 to provide a normalized measure for comparing frequencies. This method has the advantage that the denominator is stable and easily adjustable. It, thus, seems to provide an accountable measure of relative frequencies which is easily replicable and allows valid cross-corpora comparisons to be made.

Alas, the execution of this approach is not as consistent as one might hope for scholars' tabulations of word counts vary considerably. Macaulay (2005: 14, 188-

189), for example, does not include in his quantification of normalized frequency measures filled pauses such as *um* and *er*. This exclusion is problematic because it implies that filled pauses are marginal, randomly occurring linguistic elements which do not constitute an integral part of discourse grammar. Because filled pauses share functional and distributional qualities of other discourse features (Kjellmer 2003), their exclusion from word counts generated for variationist analysis is unjustified.[9]

Further, overall word counts and normalized frequency tabulations are affected by differential transcription conventions across corpora. Transcribed datasets may differ in terms of the rigour with which false starts (e.g. *th- the man*, *f- forever*) or elements such as minimal response tokens (e.g. *yeah, mhm, mm, uh-huh*) and other interjections (e.g. *right, oh*) are reproduced, and in terms of the adopted orthographic conventions. Tagliamonte (2006: 53-63) recommends the use of some idiosyncratic spellings, e.g. that cliticised auxiliaries and clitic negative particles be preceded by a space (e.g. *he 's* rather than *he's; they have nt* rather than *they haven't*) and that multi-unit discourse elements be hyphenated (e.g. *you-know* rather than *you know*). These conventions are of great value for the extraction process in variationist studies, and if they are adhered to consistently, they might not significantly affect comparisons within single datasets. However, problems arise when comparisons are made across corpora which have been transcribed using divergent conventions.

In order to illustrate the severity of these problems, Table 3 compares differential word counts and tabulation methods based on a corpus of interview data collected in north-east England (Pichler 2008). The left-hand column shows word counts which are based on a transcription of the corpus *without* hyphenation of multi-unit discourse elements; the right-hand column shows counts based on a

transcription of the same corpus *with* hyphenation of these elements. The word counts in the top rows do not include false starts, filled pauses and minimal responses; also, cliticised morphemes were counted as constituting a single unit with their preceding elements. As we move down the table, the total numbers of these elements (false starts, filled pauses, minimal responses, cliticised morphemes) are added to the total word count one at a time. A comparison of the two extremes, i.e., the top-right word count (N = 240,187) and the bottom-left word count (N = 276,707), reveals that with a medium-sized corpus, different transcription and tabulation methods can yield differences in overall word counts of over 35,000 words.

[insert Table 3]

The differential overall word counts have an unsurprising, yet profound, effect on normalized frequency counts. Based on the two extreme counts highlighted above, normalized frequency tabulations would reveal that a discourse variable which is instantiated in the corpus, say, 1,450 times occurred with a frequency of either 52.4 or 60.4 times per 10,000 words. If we were to compare these results with those obtained in the same community some 20 years earlier, where the targeted discourse variable occurred, say, 50 times per 10,000 words, we would draw very different conclusions from our results: relative stability vs. incremental increase. Whilst this is a hypothetical scenario, it illustrates well the far-reaching implications of differential word counts on the conclusions researchers draw from their analyses.

The only way forward would be for scholars to describe in minute detail, either in footnotes or appendices, how they arrived at overall word counts. Word counts would need to be broken down not just for the corpus as a whole but for individual cohorts and speakers. This would enable scholars to adjust others' word

counts and frequency tabulations to their own (and vice versa), and thus ensure the reliability of comparative results. However, it might not always be possible to provide this amount of detail about corpora. Certainly, though, collaborators on comparative projects will need to ensure that identical tabulation methods are used across all corpora investigated.

Even if tabulation methods were to be standardised across corpora, we still need to acknowledge that amount-of-speech techniques such as the one outlined above are not the most powerful tool for analysing linguistic variation and change. Frequency tabulations do not show where in the linguistic system variables occur, nor what the social and internal mechanisms are that produce variation and change in their use. A further problem is that this quantification method ignores the fact that variables do not necessarily have an equal chance of occurrence throughout an interaction but might be preferred in some stretches of discourse over others (see further Schegloff 1993; Walker 2010: 63-65). Because they reveal only general trends in the data, frequency tabulations are of little more than descriptive value in variationist discourse analysis.

4.3 Quantification of Variables without a Closed Set of Variants (2): Alternative Methods

An alternative approach to quantifying the distribution of discourse variables whose universe of variation cannot be delimited on functional or structural grounds and one which avoids some of the pitfalls of frequency tabulations is that adopted by D'Arcy (2005) in her analysis of discourse *like*. D'Arcy circumscribes the variable context for *like* according to structural criteria, i.e., as occurring in syntactically delimited positions. In contrast to the amount-of-speech approach, this method caters for the fact that *like* cannot occur with equal opportunities across the entire discourse. This approach is accountable, replicable and

generalizable. However, as D'Arcy (2005: 27) frankly acknowledges, it does not take account of the pragmatic constraints operating on the feature's distribution in discourse. Consequently, while D'Arcy's method has obvious advantages over normalized frequency tabulations, it presents only a partial, albeit in itself highly convincing, solution to the quantification conundrum in discourse variation studies. It might be interesting to explore whether it is possible to circumscribe the variable context of discourse variables according to 'discourse slots,' i.e., functional slots, as implied in Romero-Trillo (2006). However, much more work needs to be done on the functionality of discourse features before we can even begin to consider what these discourse slots may be.

5. QUALITATIVE ANALYSIS: INTEGRATING FUNCTION AS A PARAMETER IN THE ANALYSIS

As pointed out in Section 1, discourse-pragmatic features are by no means meaningless or redundant elements of discourse. A wealth of qualitative studies has shown that they are an indispensable resource for interactants in the construction and interpretation of discourse as well as in the establishment and maintenance of social rapport. These functions have been described as 'genuine grammatical functions' (Diewald 2006: 405) on the basis that grammatical function is understood as an open category which encompasses not just syntactic functions such as subject, object, complementizer, etc. but also pragmatic and procedural functions such as the signalling of epistemicity, transitions, etc. (Brinton 2006).

Some scholars have integrated quantitative with qualitative methods of data analysis and studied discourse tokens within their interactional context of occurrence in order to quantify the emergent functional categories across social

and internal factors. Adoption of this approach has revealed that consideration of discourse function is key to interpreting the usage and distribution of discourse variables: function accounts for the syntactic and interactional placement of discourse-pragmatic features (e.g. Schiffrin 1987), the creation of gendered conversational styles (e.g. Erman 1992; Holmes 1982; Macaulay 2002b), as well as variation in the formal encoding of discourse variables (e.g. Cheshire 1981; Pichler 2009; Stenström 1998). In some cases, function may even exert a more important constraint on discourse variability than social factors (e.g. Pichler 2008). In addition, close analysis of their pragmatic functions can shed light on discourse features' diachronic development (e.g. Pichler and Levey under review). A case can be made, then, for a fuller integration of qualitative methods in studies of discourse variation and change as well as a consistent inclusion of function as a factor group in quantitative analyses in order to provide accurate accounts of discourse variation and change.

Yet, despite its great hermeneutic and explanatory values, consideration of function is not an integral design feature in all discourse variation studies. Many variationists treat discourse-pragmatic features as uni-dimensional lexical elements, quantifying their distribution without any consideration of their multi-faceted pragmatic meanings (e.g. Andersen 2001; Dubois and Crouch 1975; Stubbe and Holmes 1995; Tagliamonte 2005). These studies ignore discourse variables' most fundamental property, i.e., the fact that their use is motivated solely, or at least primarily, by their functionality.

The neglect of function as a parameter in the analysis may be linked to the intrinsic multifunctionality of discourse-level features: not only do they perform different functions in different contexts of use but a single instantiation of a discourse variable can perform multiple functions simultaneously. Holmes (1984)

argues that researchers can differentiate on the basis of contextual cues the primary from the secondary function of multifunctional discourse variables and categorize them according to the former for the purpose of quantification (see also Escalera 2009; Lam 2009). This approach might seem appealing when faced with the daunting task of quantifying tokens across discourse functions. Yet it fails to reflect the fact that conversationalists may exploit discourse features' multifunctionality at strategic points in interaction (Coates 1987: 130), and that it is their multifunctional, not their unifunctional, uses that are unmarked (Cameron, McAlinden and O'Leary 1988: 77).

This is not the case with the approach adopted by Pichler (2008). When tokens perform multiple functions simultaneously, e.g. initiating or terminating a turn whilst also qualifying its content (see Extracts 1 and 2 below), Pichler categorizes these tokens as multifunctional tokens performing both functions concurrently. This taxonomy is more accurate than Holmes's since it allows researchers to reflect in quantitative terms the multifunctional nature of discourse variables. Also, it is less subjective than Holmes's taxonomy since it does not rely on researchers' intuitive judgments as to which of the multiple functions is more important in a given context.

Extract 1[10]

Barbara has just asserted that older people use more non-standard grammar than younger people.

HP: Why do you think that is.

Barbara: **I dunno**? Maybe just just just e:h education at schools.

Extract 2

HP: What accent would you say you had and do you like it?

(.)

Leah: Em. It's a mixture of probably Scottish and Geordie. But °**I dunno**°.

Yet even with this approach, an element of subjectivity remains. Analyses can be somewhat objectified by using multiple researchers to code the data independently and discuss disagreements until agreement is reached (e.g. Escalera 2009), or by involving members of the fieldwork community in the validation of functional taxonomies (e.g. Pichler 2008). Nonetheless, these procedures do not guarantee that the functional descriptions obtained across individual studies are in fact comparable. The approaches currently available for describing the pragmatic meanings and functions of discourse-pragmatic features differ in too many respects to guarantee a uniform description of their use.

Firstly, functional taxonomies of discourse-pragmatic features differ in how they account for the relationship between their various meanings. Monosemic approaches assume that a single core meaning can be isolated for individual discourse features and that variations in their use arise from their interaction with context (e.g. Dines 1980; Tsui 1991). Polysemic approaches assume that discourse features have different meanings and functions which are related through family resemblance or pragmatic extension (e.g. Buchstaller 2004; Romaine and Lange 1991).[11] Polysemic approaches are preferable for variationist discourse studies since they are better able to account for meaning variation and change (see section 3 above).

Secondly, discourse-pragmatic features can be studied within a range of different theoretical and analytical frameworks. They can be examined within a single framework, such as relevance theory (e.g. Andersen 2001; Blakemore 1988; Jucker 1993), coherence-based theory (e.g. Fraser 1996; Lenk 1998) or politeness theory (e.g. Holmes 1995). Top-down approaches such as these have

the advantage of providing unified and thus easily comparable accounts of discourse-pragmatic features. Yet, there is a risk that the focus on a single theoretical explanation for their use yields incomplete descriptions of discourse features' functional versatility (see further Aijmer 2002: 1, 8; Lam 2009: 354). Bottom-up approaches are theoretically flexible, with functional taxonomies being established through close examination of every occurrence of the targeted discourse feature in the data. They might therefore be better suited to yield comprehensive and data-driven functional taxonomies of discourse features. Consistent adoption of one of these approaches is certainly necessary to facilitate cross-corpora comparisons. Pichler (2008), for example, hypothesized that differences in the social distribution of *innit* in London English and Berwick English are caused by its performing different functions in the two dialects: in London English, where *innit* is favoured by females, it performs functions associated with a co-operative conversational style; in Berwick English, where it is favoured by males, it performs functions associated with an assertive speech style. Her attempts to test this hypothesis were hampered by the fact that the two studies on which the comparison was based (Andersen 2001; Pichler 2008) used different approaches (top-down vs. bottom-up) to describing the functionality of *innit*.

Thirdly, while there is consensus that discourse-pragmatic features function on multiple functional domains, function-based taxonomies differ in terms of the number and types of domains they identify. For example, Fischer (2000) and Schiffrin (1987) identify five domains, Bazzanella (2006) and Erman (2001) three, and Brinton (1996) only two. Some scholars prefer models such as Fischer's because it allows them 'to show the extreme functional flexibility' of discourse features (Cheshire 2007: 178). For quantitative purposes, models with

fewer domains might be preferable, though. Brinton's (1996) model allows scholars to draw as many divisions within each broad domain as are necessary for accurate qualitative descriptions of their use (e.g. signalling epistemicity or mitigation; signalling repair, turn-exchange or topic-control), but at the same time it allows them to collate these divisions into the broad domains ('interpersonal' vs. 'textual') for quantitative purposes (see further Pichler 2008, 2009). With other models, scholars might end up with too few tokens in each category to warrant statistical analysis. Collating categories (possibly on arbitrary grounds) might then be necessary to allow quantification of the data. For comparative purposes, it would in principle be possible for scholars to re-organise others' taxonomies to fit their own models. Yet this is not ideal (least of all because of the time-consuming nature of the task), and would heavily depend on scholars' provision of raw scores of tokens for each functional (sub-)category.

6. CONCLUSION

In recent decades, variationist sociolinguistics has witnessed a limited expansion in discourse variation studies. Nonetheless, we are still a very long way from having gathered the amount and quality of empirical evidence required to provide accurate descriptions of dialect variation and change in discourse, and to formulate general principles about the social and system-internal dimensions of discourse variation and change. In this paper, I have attributed the lack of progress in variationist discourse analysis to the methodological and analytical heterogeneity of the field. I have reviewed a large body of variationist discourse studies to demonstrate that the diversity of methods currently in use hampers reliable cross-corpora comparisons and consequently the formulation of generalizations about patterns of variation and change in discourse. The issue is

complicated by the fact that some of the methods currently in use yield invalid variety-specific results. To advance this line of variationist enquiry beyond its current embryonic state and to allow scholars to answer key questions about discourse variation and change, I have advocated a uniform approach to studying variation and change in this component of the grammar. It is envisaged that consistent adoption of these methods will produce the kind of valid and comparable results needed to enhance our understanding of discourse variation and change. The methods associated with this approach were presented in detail in the preceding sections. They are summarised in Table 4 which serves as a checklist for future researchers.

[insert Table 4]

In the following, I will briefly recapitulate the potential benefits of the methodology advocated here. The proposed methods for corpus description and interpretation make possible the provision of more accurate general descriptions of social and regional discourse variation than has been possible in the past. Moreover, the adoption across studies of identical quantitative and qualitative methods, as proposed here, could facilitate the achievement of three key objectives: (i) to determine where in the linguistic system intra- and inter-dialectal discourse variation occurs; (ii) to assess the extent to which correlations between social factors and discourse variation are variety-specific or pan-dialectal; and (iii) to establish what role discourse function plays in the conditioning of variation within and across dialects. To ensure that scholars compare like with like, the case has been made for conceptualising discourse features as linguistic variables.

With regard to discourse change, consistent adoption of the proposed methodology would provide scholars with the opportunity to address the following key questions: whether it is the same social groups who actuate and

advance discourse change within and across dialects; whether regionally or socially marked uses of discourse variables and variants are being levelled; and whether the grammaticalization of discourse variables and their variants progresses across dialects at similar rates and along similar syntactic and semantic-pragmatic pathways.[12]

The adoption of a uniform methodology which yields reliable and intersubjective results would allow scholars to synthesize the results from different studies into empirically grounded generalizations of discourse variation and change. Such results and generalizations are essential if our ultimate goals are to (i) develop a comprehensive account of dialect variation that encompasses all levels of the grammar; (ii) uncover potential differences in the social and internal dimensions of variation and change across different components of the grammar; and (iii) reveal potential differences in the general direction of ongoing changes across phonology, morpho-syntax and discourse. A holistic theory of language variation and change must by definition incorporate discourse variation and change, and the validity of any such theory is therefore contingent upon the generalizability of discourse variation and change studies.

The approach outlined here has potential benefits for scholars exploring how individuals deploy sociolinguistic resources to negotiate and establish social meanings and identities (Coupland 2007; Eckert 2001). Discourse variables and their variants have been shown to form part of the sociolinguistic repertoire which speakers draw on to create social identities (e.g. Mendoza-Denton 2007; Moore & Podesva 2009; Trester 2009). However, local interpretations of discourse variation rely on scholars' access to reliable quantitative survey studies which reveal the conventionalized social and functional associations of discourse variables and their variants across different varieties. Consistent adoption of the

approach to variationist discourse analysis advocated here could supply this information and thereby advance explorations of locally contextualized discourse variation.

Some of the methods advocated here may also have relevance for the exploration of how social meanings are constructed through discourse variation. The proposed methodology caters for the operation of contextual and language-internal constraints on speakers' selective use of available discourse resources, and enables scholars to explore whether the functional polysemies of available discourse resources are exploited for social meaning making, as suggested by Traugott (2001). The adoption of these methods could facilitate exploration of the social significance of discourse variation in the communities of practice framework (Eckert and McConnell-Ginet 1992) as well as in studies of 'crossing' (Rampton 1995) and 'styling' (Coupland 2007).

This paper has advocated a uniform approach to the quantitative analysis of discourse-pragmatic features which is equipped to capture the complex nature of discourse variation and change whilst also ensuring generalizability. As theoretical insights into the nature of discourse features grow and a more diverse range of features are investigated, the methodology I have proposed here will inevitably be subject to refinement. Whatever modifications may be required, the commitment to ensuring reliability and intersubjectivity is central to the variationist enterprise.

NOTES

1. An earlier version of this paper was presented at ICAME 30 (University of Lancaster, May 2009). I would like to thank the audience members for their insightful comments and questions, in particular Karin Aijmer, Gisle Andersen and Sali Tagliamonte. I am also very grateful for the detailed comments on previous written versions of the paper made by Jenny Cheshire, Stephen Levey, the anonymous reviewers as well as editor Allan Bell and associate editors David Britain and Lionel Wee, all of which have much improved the paper. Of course the ultimate responsibility for the paper has to rest with me.

2. Unlike traditional accounts of grammar, construction-based approaches do not make a distinction between 'core' and 'peripheral' phenomena (Gisborne and Trousdale 2008: 1).

3. The upsurge of interest has been most prominent in qualitative research paradigms. The pragmatic and discourse analytic research conducted into discourse-pragmatic features over the last three decades has revealed important insights into their functionality, their context sensitivity, their syntactic, semantic and prosodic integratedness as well as their evolution. In quantitative research paradigms, discourse-pragmatic features have figured far less prominently. Quantification necessarily entails a certain amount of abstraction. Nonetheless, it offers a range of complementary insights to purely qualitative studies. As I will demonstrate in this paper, quantitative studies of discourse variation and change in particular are a

valuable complement to qualitative studies because of their potential for establishing the sensitivity of discourse features to internal constraints as well as probing the robustness of their functionality across social groups, space and time.

4. Serrano (to appear) argues (very convincingly) that it is not possible to understand higher-level variability without abandoning the criterion of semantic equivalence. She argues that satisfactory explanations of syntactic variability can only be achieved by 'placing exactly in meaning differences the potential to explain variation.'

5. General extenders may contain generic lexical items only (e.g. *and that kind of stuff*) or they may contain generic along with specific lexical items (e.g. *and physical stuff like that*) (see Terraschke 2007). A structurally-based conceptualisation of discourse variables allows scholars to specify which items are included in their analyses.

6. Divergent points of departure are also common in studies of morpho-syntactic variation (Schwenter and Torres Cacoullos 2008; Hackert 2008).

7. More recently, a new interface of multivariate analysis has been developed, Rbrul, which overcomes some of the shortcomings of Goldvarb X (see Johnson 2009 for details).

8. This applies equally to phonological and morpho-syntactic variables. Variationists studying variation and change on these levels generally guard

against the problem by basing their analyses on the same number of tokens per variable per speakers (see further Guy 1974). Because of the comparative infrequency of discourse variables, this approach cannot easily be adopted in discourse variation studies and alternative measures have to be taken to guard against skewed results.

9. In Pichler's (2008) corpus of a northern English dialect, young speakers demonstrate markedly lower rates of filled pauses (8.1 per 1,000 words) than speakers from the middle and older age groups (18.2 and 15.4 per 1,000 words respectively).

10. The following transcription conventions are used:

    . falling intonation

    ? rising intonation

    (.) short pause

    : syllable lengthening

    º º reduced volume

    All informant names are pseudonyms.

11. There are also homonymic approaches which assume that the different meanings of a form are not related. However, they are not widely defended (Fischer 2006: 14).

12. Of course, the gradualness of changes associated with grammaticalization might necessitate that analyses of synchronic dialect data be supplemented with analyses of diachronic data.

REFERENCES

Aijmer, Karin. 2002. *English Discourse Particles: Evidence from a Corpus.* Amsterdam, The Netherlands: John Benjamins.

Andersen, Gisle. 2001. *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents.* Amsterdam, The Netherlands: John Benjamins.

Bailey, Guy and Jan Tillery. 2004. Some sources of divergent data in sociolinguistics. In Carmen Fought (ed.) *Sociolinguistic Variation. Critical Reflections.* Oxford, U.K.: Oxford University Press. 11–30.

Bayley, Robert. 2002. The quantitative paradigm. In J.K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds.) *The Handbook of Language Variation and Change.* Oxford, U.K.: Blackwell. 117–141.

Bazzanella, Carla. 2006. Discourse markers in Italian: Towards a 'compositional' meaning. In Kerstin Fischer (ed.) *Approaches to Discourse Particles.* Amsterdam, The Netherlands: Elsevier. 449–464.

Blakemore, Diane. 1988. *So* as a constraint on relevance. In R.M. Kempson (ed.) *Mental Representations. The Interface between Language and Reality.* Cambridge, U.K.: Cambridge University Press. 183–195.

Brinton, Laurel J. 1996. *Pragmatic Markers in English. Grammaticalization and Discourse Function.* Berlin: Mouton de Gruyter.

Brinton, Laurel J. 2006. Pathways in the development of pragmatic markers in English. In Hans van Kemenade and Bettelou Los (eds.) *The Handbook of the History of English.* Oxford, U.K.: Blackwell. 307–334.

Buchstaller, Isabelle. 2004. The sociolinguistics constraints on the quotative system - British English and US English compared. Unpublished PhD dissertation, University of Edinburgh, Edinburgh, U.K.

Buchstaller, Isabelle. 2006. Diagnostics of age-graded linguistic behaviour: The case of the quotative system. *Journal of Sociolinguistics* 10: 3–30.

Buchstaller, Isabelle. 2009. The quantitative analysis of morpho-syntactic variation: Constructing and quantifying the denominator. *Language and Linguistics Compass* 3: 1010–1033.

Buchstaller, Isabelle and Alexandra D'Arcy. 2009. Localized globalization: A multi-local, multivariate investigation of quotative *be like*. *Journal of Sociolinguistics* 13: 291–331.

Cameron, Deborah, Fiona McAlinden and Kathy O'Leary. 1988. Lakoff in context: The social and linguistic functions of tag questions. In Jennifer Coates and Deborah Tannens (eds.) *Women in Their Speech Communities. New Perspectives on Language and Sex.* Longman, U.K.: London. 74–93.

Cheshire, Jenny. 1981. Variation in the use of *ain't* in an urban British English dialect. *Language in Society* 10: 365–381.

Cheshire, Jenny. 1987. Syntactic variation, the linguistic variable, and sociolinguistic theory. *Linguistics*: 25: 257–282.

Cheshire, Jenny. 2007. Discourse variation, grammaticalisation and stuff like that. *Journal of Sociolinguistics* 11: 155–193.

Cheshire, Jenny, Paul Kerswill and Anne Williams. 2005. Phonology, grammar and discourse in dialect convergence. In Peter Auer, Frans Hinskens and Paul Kerswill (eds.) *Dialect Change. Convergence and Divergence in European Languages.* Cambridge, U.K.: Cambridge University Press. 135–167.

Coates, Jennifer. 1987. Epistemic modality and spoken discourse. *Transactions of the Philological Society* 85: 110–131.

Coupland, Nikolas. 2007. *Style. Language Variation and Identity.* Cambridge, U.K.: Cambridge University Press.

D'Arcy, Alexandra. 2005. Like: Syntax and development. Unpublished PhD dissertation, University of Toronto, Toronto, Canada.

Diewald, Gabriele. 2006. Discourse particles and modal particles as grammatical elements. In Kerstin Fischer (ed.) *Approaches to Discourse Particles.* Amsterdam, The Netherlands: Elsevier. 406–426.

Dines, Elizabeth. 1980. Variation in discourse – 'and stuff like that'. *Language in Society* 9: 13–31.

Dubois, Betty L. and Isabel Crouch. 1975. The question of tag questions in women's speech: They don't really use more of them, do they? Language in Society 4: 289–294.

Dubois, Sylvie. 1992. Extension particles, etc. *Language Variation and Change* 4: 179–203.

Eckert, Penelope. 2001. Style and social meaning. In Penelope Eckert and John R. Rickford (eds.) *Style and Sociolinguistic Variation.* Cambridge, U.K.: Cambridge University Press. 119–126.

Eckert, Penelope and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21: 461–490.

Erman, Britt. 1992. Female and male usage of pragmatic expressions in same-gender and mixed-gender interaction. *Language Variation and Change* 4: 217–234.

Erman, Britt. 2001. Pragmatic markers revisited with a focus on *you know* in adult and adolescent talk. *Journal of Pragmatics* 32: 1337–1359.

Escalera, Elena A. 2009. Gender differences in children's use of discourse markers: Separate worlds or different contexts? *Journal of Pragmatics* 41: 1887–1907.

Ferrara, Kathleen W. 1997. Form and function of the discourse marker *anyway*: Implications for discourse analysis. *Linguistics* 35: 343–378.

Ferrara, Kathleen and Barbara Bell. 1995. Sociolinguistic variation and discourse function of constructed dialogue of introducers: The case of *be + like*. *American Speech* 70: 265–290.

Fischer, Kerstin. 2000. Discourse particles, turn-taking, and the semantics-pragmatics interface. *Revue de Sémantique et Pragmatique* 8*:* 111–137.

Fischer, Kerstin. 2006. Towards an understanding of the spectrum of approaches to discourse particles: Introduction to the volume. In Kerstin Fischer (ed.) *Approaches to Discourse Particles.* Amsterdam, The Netherlands: Elsevier. 1–20.

Fraser, Bruce.1996. Pragmatic markers. *Pragmatics* 6: 167–190.

Freed, Alice F. and Alice Greenwood. 1996. Women, men, and type of talk: What makes the difference. *Language in Society* 25: 1–26.

Fuller, Janet M. 2003. The influence of speaker roles on discourse marker use. *Journal of Pragmatics* 35: 23–45.

Gisborne, Nikolas and Graeme Trousdale. 2008. Constructional approaches to language-particular description. In Graeme Trousdale and Nikolas Grisborne (eds.) *Constructional Approaches to English Grammar.* Berlin: Mouton de Gruyter. 1–4.

Guy, Gregory R. 1974. Variation in the group and the individual: The case of final stop deletion. *Pennsylvania Working Papers on Linguistic Change and Variation* 1: 1–75.

Hackert, Stephanie. 2008. Counting and coding the past: Circumscribing the variable context in quantitative analyses of past inflection. *Language Variation and Change* 20: 127–153.

Holmes, Janet. 1982. The functions of tag questions. *English Language Research Journal* 4: 40-65.

Holmes, Janet. 1984. Hedging your bets and sitting on the fence: Some evidence for hedges as support structures. *Te Reo* 27: 47–62.

Holmes, Janet. 1995. *Women, Men and Politeness.* London, U.K.: Longman.

Ito, Rika and Sali Tagliamonte. 2003. *Well* weird, *right* dodgy, *very* strange, *really* cool: Layering and recycling in English intensifiers. *Language in Society* 32: 257–279.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard. Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–383.

Jucker, Andreas H. 1993. The discourse marker *well*: A relevance-theoretical account. *Journal of Pragmatics* 19: 435–452.

Jucker, Andreas H. and Sarah W. Smith. 1998. *And people just you know like 'wow'*: Discourse markers as negotiating strategies. In Andreas H. Jucker and Yul Ziv (eds.) *Discourse Markers. Descriptions and Theory.* Amsterdam, The Netherlands: John Benjamin. 171–201.

Kallen, Jeffrey and John Kirk. 2007. ICE-Ireland: Local variations on global standards. In Joan C. Beal, Karen P. Corrigan and Hermann L. Moisl (eds.) *Creating and Digitizing Corpora. Vol. 1: Synchronic Databases.* Basingstoke, U.K.: Palgrave Macmillan. 121–162.

Kjellmer, Göran. 2003. Hesitation. In defence of *er* and *erm*. *English Studies* 2: 170–198.

Kyratzis, Amy and Susan Ervin-Tripp. 1999. The development of discourse markers in peer interaction. *Journal of Pragmatics* 31: 1321–1338.

Labov, William. 1963. The social motivation of a sound change. *Word* 19: 273–

309.

Labov, William. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Centre for Applied Linguistics.

Labov, William. 1972. *Sociolinguistic Patterns.* Oxford, U.K.: Blackwell.

Labov, William. 1998. The intersection of sex and social class in the course of linguistic change. In Jenny Cheshire and Peter Trudgill (eds.) *The Sociolinguistics Reader. Vol. 2: Gender and Discourse*. Oxford, U.K.: Oxford University Press. 7–52.

Lam, Phoenix W.Y. 2009. The effect of text type on the use of *so* as a discourse particle. *Discourse Studies* 11: 353–372.

Lavandera, Beatriz R. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7: 171–182.

Lenk, Uta. 1998. *Marking Discourse Coherence: Functions of Discourse Markers in Spoken English.* Tübingen: Gunter Narr.

Macaulay, Ronald. 2001. *You're like 'why not?'* The quotative expressions of Glasgow adolescents. *Journal of Sociolinguistics* 5: 3–21.

Macaulay, Ronald. 2002a. Discourse variation. In J.K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds.) *The Handbook of Language Variation and Change*. Oxford, U.K.: Blackwell. 283–305.

Macaulay, Ronald. 2002b. Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. *Journal of Sociolinguistics* 6: 398–417.

Macaulay, Ronald. 2005. *Talk That Counts. Age, Gender, and Social Class Differences in Discourse.* Oxford, U.K.: Oxford University Press.

Macaulay, Ronald. 2006. Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change* 18: 267–283.

Mauranen, Anna. 2004. ‘They’re a little bit different’ ...: Observations on hedges in academic talk. In Karin Aijmer (ed.) *Patterns in Spoken and Written Corpora.* Amsterdam, The Netherlands: John Benjamins. 173–197.

Mendoza-Denton, Norma. 2007. *Homegirls. Language and Cultural Practice among Latina Youth Gangs.* Oxford, U.K.: Wiley-Blackwell.

Meyerhoff, Miriam. 1994. Sounds pretty ethnic, eh?: A pragmatic particle in New Zealand English. *Language in Society* 23: 367–388.

Moore, Emma and Robert Podesva. 2009. Style, indexicality, and the social meaning of tag questions. *Language in Society* 38: 447–485.

Overstreet, Maryann. 1999. *Whales, Candlelight, and stuff like that.* Oxford, U.K.: Oxford University Press.

Pichler, Heike. 2008. A qualitative-quantitative study of negative auxiliaries in a northern English dialect: I DON’T KNOW and I DON’T THINK**,** *innit*? Unpublished PhD dissertation, University of Aberdeen, Aberdeen, U.K.

Pichler, Heike. 2009. The functional and social reality of discourse variants in a northern English dialect: I DON’T KNOW and I DON’T THINK compared. *Intercultural Pragmatics* 6: 561–596.

Pichler, Heike and Stephen Levey. under review. Variationist theory meets grammaticalization theory: General extenders in north-east England *and that*.

Preston, Dennis R. 1991. Sorting out the variables in sociolinguistic theory. *American Speech* 66: 33–56.

Rampton, Ben. 1995. *Crossing.* London, U.K.: Longman.

Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14: 367–381.

Rickford, John, Isabelle Buchstaller, Thomas Warsow, Arnold Zwicky, Elizabeth Closs Traugott. 2007. Intensive and quotative *all*: Something old, something new. *American Speech* 82: 3–31.

Romaine, Suzanne. 1984. On the problem of syntactic variation and pragmatic meaning in sociolinguistic theory. *Folia Linguistica* 18: 409–437.

Romaine, Suzanne and Deborah Lange. 1991. The use of *like* as a marker of reported speech and thought: A case of grammaticalization in progress. *American Speech* 66: 227–279.

Romero-Trillo, Jesus. 2006. Discourse markers. In Keith Brown (ed.) *Encyclopedia of Language and Linguistics.* 2nd ed. Amsterdam: Elsevier. 639–642.

Sankoff, Gillian. 1973. Above and beyond phonology in variable rules. In Charles-James N. Bailey and Roger W. Shuy (eds.) *New Ways of Analysing Variation in English.* Washington: Georgetown University Press. 44–61.

Sankoff, David, Sali Tagliamonte and E. Smith. 2005. Goldvarb X. A multivariate analysis application. Department of Linguistics, University of Toronto, and Department of Mathematics, University of Ottawa.

Schegloff, Emanual. 1993. Reflections on quantification in the study of conversation. *Research on Language and Social Interaction* 26: 99–128.

Schiffrin, Deborah. 1987. *Discourse Markers.* Cambridge, U.K.: Cambridge University Press.

Schleef, Erik. 2008. The 'lecturer's *ok*' revisited: Changing discourse conventions and the influence of academic division. *American Speech* 83: 62-84.

Schwenter, Scott and Rena Torres Cacoullos. 2008. Defaults and indeterminacy in temporal grammaticalization: The 'perfect' road to perfectives. *Language Variation and Change* 20: 1–39.

Serrano, Maria José. to appear. Morphosyntactic variation in Spain. In Manuel Diáz Campos (ed.) *The Handbook of Spanish Linguistics.* Oxford, U.K.: Blackwell.

Siegel, Muffy. 2002. 'Like': The discourse particle and semantics. Journal of Semantics 19: 35–71.

Stenström, Anna-Brita. 1998. From sentence to discourse: *Cos (because)* in teenage talk. In Andres H. Jucker and Yael Ziv (eds.) *Discourse Markers. Descriptions and Theory.* Amsterdam, The Netherlands: John Benjamins. 127–146.

Stubbe, Maria and Janet Holmes. 1995. *You know, eh* and other exasperating expressions: An analysis of social and stylistic variation in the use of pragmatic particles in a sample of New Zealand English. *Language and Communication* 15: 63–88.

Tagliamonte, Sali. 2002. Comparative sociolinguistics. In J.K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.) *The Handbook of Language Variation and Change.* Oxford, U.K.: Blackwell. 729–763.

Tagliamonte, Sali. 2005. So who? Like how? Just what? Discourse markers in the conversations of young Canadians. *Journal of Pragmatics* 37: 1896–1915.

Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation.* Cambridge, U.K.: Cambridge University Press.

Tagliamonte, Sali A. and Derek Denis. 2010. The *stuff* of change: General extenders in Toronto, Canada. *English Language and Linguistics* 38: 1-34.

Tagliamonte, Sali and Rachel Hudson. 1999. Be like et al. beyond America: The quotative system in British and Canadian youth. *Journal of Sociolinguistics* 3: 147–172.

Terraschke, Agnes. 2007. Use of general extenders by German non-native

speakers of English. *International Review of Applied Linguistics in Language Teaching* 45: 141–160.

Torres Cacoullos, Rena. 2001. From lexical to grammatical to social meaning. *Language in Society* 30: 443–478.

Traugott, Elizabeth C. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper presented at ICHL XII, 17 August, Manchester. http://www.stanford.edu/~traugott/papers/discourse.pdf

Traugott, Elizabeth Closs. 2001. Zeroing in on multifunctionality and style. In Penelope Eckert and John R. Rickford (eds.) *Style and Sociolinguistic Variation*. Cambridge, U.K.: Cambridge University Press. 127–S136.

Trester, Anna Marie. 2009. Discourse marker 'oh' as a means for realizing the identity potential of constructed dialogue in interaction. *Journal of Sociolinguistics* 13: 147–168.

Tsui, Amy B. M. 1991. The pragmatic functions of *I don't know*. *Text* 11: 607–622.

Verdonik, Darink, Andrej Žgank and Agnes Pisanski Peterlin. 2009. The impact of context on discourse marker use in two conversational genres. *Discourse Studies* 10: 759–775.

Vincent, Diane and David Sankoff. 1993. Punctors. A pragmatic variable. *Language Variation and Change* 4: 205–216.

Walker, James A. 2010. *Variation in Linguistic Systems*. London, U.K.: Routledge.

Waters, Cathleen. 2009. Transatlantic divergence: The social and linguistic correlates of *actually/really* variation. Paper presented NWAV 38, 24 October, Ottawa.

Address correspondence to:

Heike Pichler

School of Languages

University of Salford

Salford, Greater Manchester

M5 4WT

U.K.

Phone: +44 (0)161 295 4575

Fax: +44 (0)161 295 4646

e-mail: h.pichler@salford.ac.uk

Table 1. Effects of contextual factors on the use of discourse-pragmatic features (unless otherwise stated, studies are based American English data)

	data	factors constraining the variation	impact on
Cameron, McAlinden and O'Leary (1988)	British English same- and mixed-sex conversations; broadcast talk	- speaker roles and relationships	function
Escalera (2009)	naturalistic peer conversations between 3- to 5-year-olds	- activity context	function
Freed and Greenwood (1996)	same-sex dyadic conversations	- discourse type - topic - goal of interaction	frequency
Fuller (2003)	interviews with strangers; casual conversations between friends	- discourse type - speaker roles and relationships	frequency
Jucker and Smith (1998)	conversations between pairs of students (friends and strangers)	- speaker roles and relationships	frequency
Kyratzis and Ervin-Tripp	conversations between 4- and 7-year old best-	- activity context	function

(1999)	friend dyads		
Lam (2009)	spoken corpus of Hong Kong English	- discourse type	frequency function
Mauranen (2004)	spoken academic discourse	- discourse type	function
Redeker (1990)	conversations between friends and strangers	- speaker roles and relationships	function
Schleef (2008)	university lectures	- discourse type - topic	frequency
Stubbe and Holmes (1995)	corpus of spoken New Zealand English	- discourse type - formality	frequency
Verdonik, Žgank and Pisanski Peterlin (2009)	Slovenian telephone conversations and television interviews	- discourse type - topic - speaker roles and relationships - communication channel - attitudes towards interaction	frequency

Table 2. Textual metadata required for reliable cross-corpora comparisons

	textual metadata	examples
social factors	number of participants	
	relationship between participants	intimate/casual; symmetrical/asymmetrical
	role of participants	peers; interviewer/interviewee
	participants' assumed shared knowledge	given/to be negotiated
physical factors	channel of communication	present/distant
	place of recording	private/public
stylistic factors	formality	formal/informal
semantic factors	topics discussed	general/specialized prepared/non-prepared
psychological factors	attitudes towards interaction	engaged/withdrawn
	attitudes towards topic	objective/subjective
	goals of interaction	phatic/informational
discourse factors	speech situation/event	monologic/dialogic/multilogic one-to-one/one-to-many spontaneous/non-spontaneous structured/unstructured
	activity context	discussion/play/narrative

Table 3. Results of differential frequency tabulations

	total N *without* hyphenation	total N *with* hyphenation
total N of words	250,682	**240,187**
+ false starts (N = 2,962)	253,644	243,149
+ filled pauses (N = 3,884)	257,528	247,033
+ minimal responses (N = 2,402)	259,930	249,435
+ cliticised morphemes (N = 16,777)	**276,707**	266,212

Table 4. Summary of methods advocated for a uniform discourse variation analysis

	To ensure reliability, generalizability and comparability, scholars need to:
corpus construction	# provide detailed textual metadata about their corpora # be familiar with and consider potential effects of contextual constraints on the observed variation
discourse variables	# close the set of possible variants wherever possible # state clearly on what grounds the variable context has been delimited (e.g. functional comparability or structural similarity) # ensure that the variable context is delimited in accordance with their goals of investigating variation and change in linguistic form, pragmatic function or both # be mindful of the methodological and theoretical limitations of functionally-based circumscriptions of the variable context
data quantification	# conduct multivariate analyses wherever possible # take into account the limitations of analyses based on normalized frequency tabulations, and detail which linguistic elements are included in generating such tabulations # consider delimiting the variable context on syntactic or functional grounds in the absence of alternative delimitation procedures

qualitative analysis	# conduct qualitative analyses of discourse variables in order to include function as a factor group in the analysis # cater for the multifunctionality and polysemic nature of discourse variables # take a bottom-up approach to establishing functional taxonomies of discourse variables # categorise tokens in ways that allow maximum flexibility in data quantification